Linear Regression

For (X, Y) a pair of random variables with values in $\mathbb{R}^p \times \mathbb{R}$ we assume that

$$E(Y|X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j = (1, X^T)\beta$$

with $\beta \in \mathbb{R}^{p+1}$.

This "model" of the conditional expectation is *linear in the parameters.*

The *predictor function* for a given β is

$$f_\beta(x) = (1, x^T)\beta.$$

A more practical and relaxed attitude towards linear regression is to say that

$$E(Y|X) \simeq \beta_0 + \sum_{j=1}^{p} X_j \beta_j = (1, X^T)\beta$$

where the precision of the approximation of the conditional mean by a linear function is swept under the carpet. All mathematical derivations rely on assuming that the conditional mean is exactly linear, but in reality we will almost always regard linear regression as an approximation. Linearity is for differentiable functions a good local approximation and it may extend reasonably to the convex hull of the x_i's. But we must make an attempt to check that by some sort of model control. Having done so, interpolation – prediction for a new x in the convex hull of the observed x_i's – is usually OK, whereas extrapolation is always dangerous. Extrapolation is strongly model dependent and we do not have data to justify that the model is adequate for extrapolation.

Least Squares

With $\mathbf{X}$ the $N \times (p+1)$ data matrix including the column $\mathbf{1}$ the column of predicted values for given β is $\mathbf{X}\beta$.

The *residual sum of squares* is

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - (1, x_i^T)\beta)^2 = ||\mathbf{y} - \mathbf{X}\beta||^2.$$

The *least squares estimate* of β is

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \mathrm{RSS}(\beta).$$

Figure 3.1 – Geometry

The linear regression seeks a p-dimensional, affine representation – a hyperplane – of the $p+1$-dimensional variable (X, Y).

The direction of the Y-variable plays a distinctive role – the error of the approximating hyperplane is measured parallel to this axis.

The Solution – the Calculus Way

Since $\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$

$$D_\beta \mathrm{RSS}(\beta) = -2(\mathbf{y} - \mathbf{X}\beta)^T\mathbf{X}$$

The derivative is a $1 \times p$ dimensional matrix – a *row vector*. The *gradient* is $\nabla_\beta \mathrm{RSS}(\beta) = D_\beta \mathrm{RSS}(\beta)^T$.

$$D_\beta^2 \mathrm{RSS}(\beta) = 2\mathbf{X}^T\mathbf{X}.$$

If $\mathbf{X}$ has rank $p+1$, $D_\beta^2 \mathrm{RSS}(\beta)$ is (globally) positive definite and there is a unique minimizer found by solving $D_\beta \mathrm{RSS}(\beta) = 0$. The solution is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

For the differentiation it may be useful to think of RSS as a composition of the function $a(\beta) = (\mathbf{y} - \mathbf{X}\beta)$ from $\mathbb{R}^p$ to $\mathbb{R}^N$ with derivative $D_\beta a(\beta) = -\mathbf{X}$ and then the function $b(z) = ||z||^2 = z^T z$ from $\mathbb{R}^N$ to $\mathbb{R}$ with derivative $D_z b(z) = 2z^T$. Then by the chain rule

$$D_\beta \mathrm{RSS}(\beta) = D_z b(a(\beta)))D_\beta a(\beta) = -2(\mathbf{y} - \mathbf{X}\beta)^T\mathbf{X}$$

The Solution – the Geometric Way (Figure 3.2)

With $V = \{\mathbf{X}\beta \mid \beta \in \mathbb{R}^p\}$ the column space of $\mathbf{X}$ the quantity

$$\mathrm{RSS}(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2$$

is minimized whenever $\mathbf{X}\beta$ is the orthogonal projection of $\mathbf{y}$ onto V. The column space projection equals

$$P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

whenever $\mathbf{X}$ has full rank $p+1$.

In this case $\mathbf{X}\beta = P\mathbf{y}$ has the unique solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

One verifies that P is in fact the projection by verifying three characterizing properties:

$$\begin{aligned} PV &= V \\ P^2 &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = P \\ P^T &= (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = P \end{aligned}$$

If $\mathbf{X}$ does not have full rank p the projection is still well defined and can be written as

$$P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$$

where $(\mathbf{X}^T\mathbf{X})^-$ denotes a generalized inverse. A generalized inverse of a matrix A is a matrix A^- with the property that

$$AA^-A = A$$

and using this property one easily verifies the same three conditions for the projection. In this case, however, there is not a unique solution to $\mathbf{X}\beta = P\mathbf{y}$.

Distributional Results – Conditionally on X

$$\varepsilon_i = Y_i - (1, X_i)^T\beta$$

Assumption 1: $\varepsilon_1, \ldots, \varepsilon_N$ conditionally on $X_1, \ldots, X_N$ are uncorrelated with mean value 0 and same variance σ^2.

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(Y_i - \mathbf{X}\hat{\beta})^2 = \frac{1}{N-p-1}||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2 = \frac{\text{RSS}(\hat{\beta})}{N-p-1}$$

Then $V(\mathbf{Y}|\mathbf{X}) = \sigma^2 I_N$

$$\begin{aligned} E(\hat{\beta}|\mathbf{X}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta \\ V(\hat{\beta}|\mathbf{X}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2 I_N\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\ E(\hat{\sigma}^2|\mathbf{X}) &= \sigma^2 \end{aligned}$$

Distributional Results – Conditionally on X

Assumption 2: $\varepsilon_1, \ldots, \varepsilon_N$ conditionally on $X_1, \ldots, X_N$ are i.i.d. $N(0, \sigma^2)$.

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$
$$(N-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p-1}.$$

The standardized Z-score

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{N-p-1}.$$

Or more generally for any $a \in \mathbb{R}^{p+1}$

$$\frac{a^T\hat{\beta} - a^T\beta}{\hat{\sigma}\sqrt{a^T(\mathbf{X}^T\mathbf{X})^{-1}a}} \sim t_{N-p-1}.$$

Minimal Variance, Unbiased Estimators

Does there exist a minimal variance, unbiased estimator of β? We consider *linear estimators only*

$$\tilde{\beta} = C^T\mathbf{Y}$$

for some $N \times p$ matrix C requiring that

$$\beta = C^T\mathbf{X}\beta$$

for all β. That is, $C^T\mathbf{X} = I_{p+1} = \mathbf{X}^TC$. Under *Assumption 1*

$$V(\tilde{\beta}|\mathbf{X}) = \sigma^2 C^T C,$$

and we have

$$\begin{aligned} V(\hat{\beta} - \tilde{\beta}|\mathbf{X}) &= V(((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - C^T)Y|\mathbf{X}) \\ &= \sigma^2((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - C^T)((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - C^T)^T \\ &= \sigma^2(C^TC - (\mathbf{X}^T\mathbf{X})^{-1}) \end{aligned}$$

The matrix $C^TC - (\mathbf{X}^T\mathbf{X})^{-1}$ is positive semidefinite, i.e. for any $a \in \mathbb{R}^{p+1}$

$$a^T(\mathbf{X}^T\mathbf{X})^{-1}a \leq a^TC^TCa$$

Gauss-Markov's Theorem

Theorem 1. *Under Assumption 1 the least squares estimator of β has minimal variance among all linear, unbiased estimators of β.*

This means that for any $a \in \mathbb{R}^p$, $a^T\hat{\beta}$ has minimal variance among all estimators of $a^T\beta$ of the form $a^T\tilde{\beta}$ where $\tilde{\beta}$ is a linear, unbiased estimator.

It also means that $V(\tilde{\beta}) - (\mathbf{X}^T\mathbf{X})^{-1}$ is positive semidefinite – or in the partial ordering on positive semidefinite matrices

$$V(\tilde{\beta}) \succeq (\mathbf{X}^T\mathbf{X})^{-1}.$$

Why look any further – we have found the optimal estimator....?

Biased Estimators

The *mean squared error* is

$$\text{MSE}_\beta(\tilde{\beta}) = E_\beta(||\tilde{\beta} - \beta||^2).$$

By Gauss-Markov's Theorem $\hat{\beta}$ is optimal for all β among the *linear, unbiased* estimators.

Allowing for biased – possibly linear – estimators we can achieve improvements of the MSE for some β – perhaps at the expense of some other β.

The *Stein estimator* is a *non-linear, biased* estimator, which under *Assumption 2* has *uniformly* smaller MSE than $\hat{\beta}$ whenever $p \geq 3$.

You can find more information on the Stein estimator, discovered by James Stein, in the Wikipedia article. The result was not the expectation of the time. In the Gaussian case it can be hard to imagine that there is an estimator that in terms of MSE performs uniformly better than $\hat{\beta}$. Digging into the result it turns out to be a consequence of the geometry in $\mathbb{R}^p$ in dimensions greater than 3. In terms of MSE the estimator $\hat{\beta}$ is *inadmissible*.

The consequences that people have drawn of the result has somewhat divided the statisticians. Some has seen it as the ultimate argument for the non-sense estimators we can get as optimal or at least admissible estimators. If $\hat{\beta}$ – the MLE in the Gaussian setup – is not admissible there is something wrong with the concept of admissibility. Another reaction is that the Stein estimator illustrates that MLE (and perhaps unbiasedness) is problematic for small sample sizes. To get better performing estimators we should consider biased estimators. However, there are some rather arbitrary choices in the formulation of the Stein estimator, which makes it hard to accept it as an estimator we would use in practice.

We will in the course consider a number of biased estimators. However, the reason is not so much due to the Stein result. The reason is much more a consequence of a favorable bias-variance tradeoff when p is large compared to N, which can improve prediction accuracy.

Shrinkage Estimators

If $\tilde{\beta} = C^T Y$ is some, biased, linear estimator of β we define the estimator

$$\tilde{\beta}_\gamma = \gamma\hat{\beta} + (1-\gamma)\tilde{\beta}, \quad \gamma \in [0,1].$$

It is biased. The mean squared error is a quadratic function in γ, and the optimal shrinkage parameter $\gamma(\beta)$ can be found – it depends upon β! Using the *plug-in principle* we get an estimator

$$\tilde{\beta}_{\gamma(\hat{\beta})} = \gamma(\hat{\beta})\hat{\beta} + (1-\gamma(\hat{\beta}))\tilde{\beta}.$$

It *could* be uniformly better – but the point is that for β where $\tilde{\beta}$ is not *too biased* it can be a substantial improvement over $\hat{\beta}$.

Take home message: Bias is a way to introduce *soft model restrictions* with a *locally* – not globally – favorable bias-variance tradeoff.

There are other methods than the plug-in principle, which can be useful. It depends on the concrete biased estimator whether we can fit a useful, closed form expression for $\gamma(\beta)$ or whether we will need some additional estimation techniques.

Regression in practice

How does the computer do multiple linear regression? Does it compute the matrix $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$?

NO!

If the columns $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are orthogonal the solution is

$$\hat{\beta}_i = \frac{<\mathbf{y}, \mathbf{x}_i>}{||\mathbf{x}_i||^2}$$

Here either $\mathbf{1}$ is included and orthogonal to the other $\mathbf{x}_i$'s or all variables have first been centered.

We use here the inner product notation $< \mathbf{y}, \mathbf{x}_i >$ as the book, which in terms of matrix multiplication could just be written $\mathbf{y}^T\mathbf{x}_i$. The norm is also given in terms of the inner product, $||\mathbf{x}_i||^2 =< \mathbf{x}_i, \mathbf{x}_i >= \mathbf{x}_i^T\mathbf{x}_i$.

Orthogonalization

If $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are not orthogonal the *Gram-Schmidt* orthogonalization produces an orthogonal basis $\mathbf{z}_1, \ldots, \mathbf{z}_p$ spanning the same column space (Algorithm 3.1). Thus $\hat{\mathbf{y}} = \mathbf{X}^T\hat{\beta} = \mathbf{Z}^T\bar{\beta}$ with

$$\bar{\beta}_i = \frac{< \mathbf{y}, \mathbf{z}_i >}{||\mathbf{z}_i||^2}.$$

By Gram-Schmidt

- $\text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_i\} = \text{span}\{\mathbf{z}_1, \ldots, \mathbf{z}_i\}$, hence $\mathbf{z}_p \perp \mathbf{x}_1, \ldots, \mathbf{x}_{p-1}$
- $\mathbf{x}_p = \mathbf{z}_p + \mathbf{w}$ with $\mathbf{w} \perp \mathbf{z}_p$.

Hence

$$\hat{\beta}_p = \mathbf{z}_p^T\mathbf{x}_p\hat{\beta}_j = \mathbf{z}_p^T\mathbf{X}^T\hat{\beta} = \mathbf{z}_p^T\mathbf{Z}^T\bar{\beta} = \bar{\beta}_p.$$

It is assumed above that the p column vectors are linearly independent and thus span a space of dimension p. Equivalently the matrix $\mathbf{X}$ has rank p, which in particular requires that $p \leq N$. If we have centered the variables first we need $p \leq N+1$. Since we can permute the columns into any order prior to this argument, the conclusion is not special to $\hat{\beta}_p$. Any of the coefficients can be seen as the residual effect of $\mathbf{x}_i$ when we correct for all the other variables.

Figure 3.4 – Gram-Schmidt

By Gram-Schmidt the multiple regression coefficient $\hat{\beta}_p$ equals the coefficient for $\mathbf{z}_p$.

If $||\mathbf{z}_p||^2$ is small the variance

$$V(\hat{\beta}_p) = \frac{\sigma^2}{||\mathbf{z}_p||^2}$$

is large and the estimate is uncertain.

For observational $\mathbf{x}_i$'s this problem occurs for highly correlated observables.

The QR-decomposition

The matrix version of Gram-Schmidt is the decomposition

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma}$$

where the columns in $\mathbf{Z}$ are orthogonal and the matrix $\boldsymbol{\Gamma}$ is upper triangular. If

$$\mathbf{D} = \text{diag}(||\mathbf{z}_1||, \ldots, ||\mathbf{z}_p||)$$

$$\begin{aligned}\mathbf{X} &= \underbrace{\mathbf{Z}\mathbf{D}^{-1}}_{\mathbf{Q}}\underbrace{\mathbf{D}\boldsymbol{\Gamma}}_{\mathbf{R}} \\ &= \mathbf{Q}\mathbf{R}\end{aligned}$$

This is the *QR-decomposition* with $\mathbf{Q}$ an *orthogonal matrix* and $\mathbf{R}$ *upper triangular*.

Using the QR-decomposition for Estimation

If $\mathbf{X} = \mathbf{Q}\mathbf{R}$ is the QR-decomposition we get that

$$\begin{aligned}\hat{\beta} &= (\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\ &= \mathbf{R}^{-1}(\mathbf{R}^T)^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\ &= \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}.\end{aligned}$$

Or we can write that $\hat{\beta}$ is the solution of

$$\mathbf{R}\hat{\beta} = \mathbf{Q}^T\mathbf{y},$$

which is easy to solve as $\mathbf{R}$ is upper triangular.

We also get

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{Q}\mathbf{Q}^T\mathbf{y}.$$